# [https://avatars2.githubusercontent.com/u/4156894?v=3&s=100](http://www.calstatela.edu/centers/hipic) CIS3200-02 Term Project Tutorial

#### Authors: Samuel Mendoza;Israel Alegria, Hector Pedro-Juan,

#### Julio Montiel Nava, Marco Rodriguez

#### Instructor: [Jongwook Woo](https://www.linkedin.com/in/jongwook-woo-7081a85)

#### Date: 05/17/2020

**Microsoft Azure Machine Learning**

**TEAM 2**

05/17/2020

**Data Processing and Analytics COVID-19 Analysis**

**Objectives**

**List what your objectives are.** In this hands-on lab, you will learn how to:

* Prepared the Data
* Execute R Script
* Clean the Data
* SQL Transformation
* Training Clustering Model

Step 1: Preparing the Data

**This is to manually retrieve the files and data necessary to mimic our Azure ML portion of the project.**

<https://www.dropbox.com/s/smw5aein7i8zr1d/covid-19-data-resource-hub-covid-19-case-counts%20%281%29.zip?dl=0>

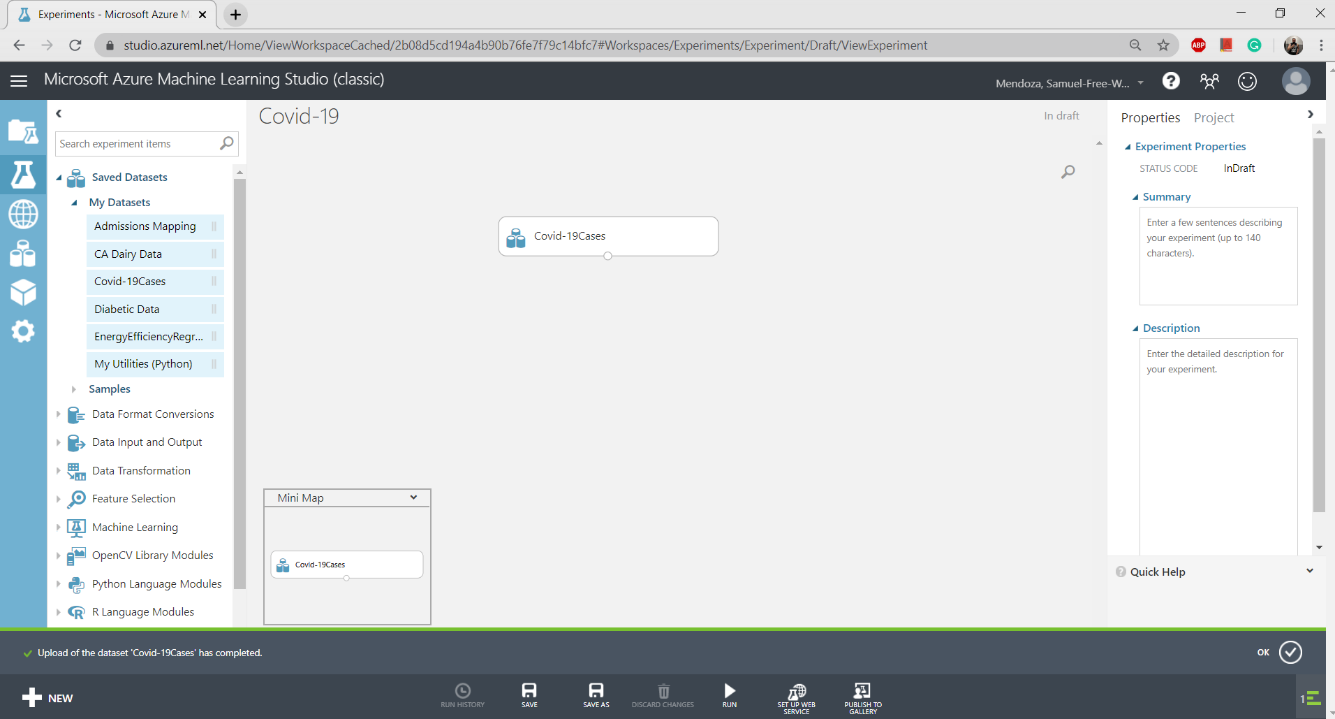
https://github.com/juliom12/covid\_19\_cases

1. Open and Login into Azure ML (<https://studio.azureml.net/>)
2. To upload a database, go to Experiments, click on +NEW, click on Databases, and upload From Local File.
3. Create a new dataset name Covid-19 Cases by uploading Covid\_19\_combined.csv from the extracted folder files.

Step 2: Create Experiment and Add Modules

**This step is to import the database and establish a new experiment canvas in AzureML.**

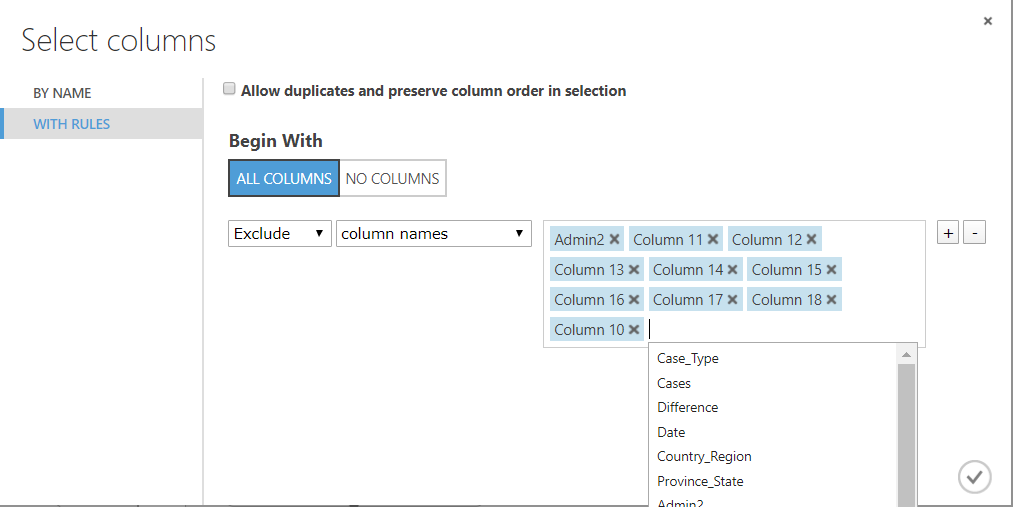
1. In the Studio, click on the bottom left (**+NEW**), Experiment 🡪 Blank Experiment
2. To upload a database, go to Experiments, click on +NEW, click on Databases, and upload From Local File.
3. Create a new dataset name **Covid-19 Cases** by uploading **Covid\_19\_combined.csv** from the extracted folder files.
4. Change the title of experiment to Covid-19
5. On the left side of the experiments pane, expand **Saved Datasets,** expand **My Datasets,** and click and drag Covid-19 Cases to the canvas.



Step 3: Excluding Columns

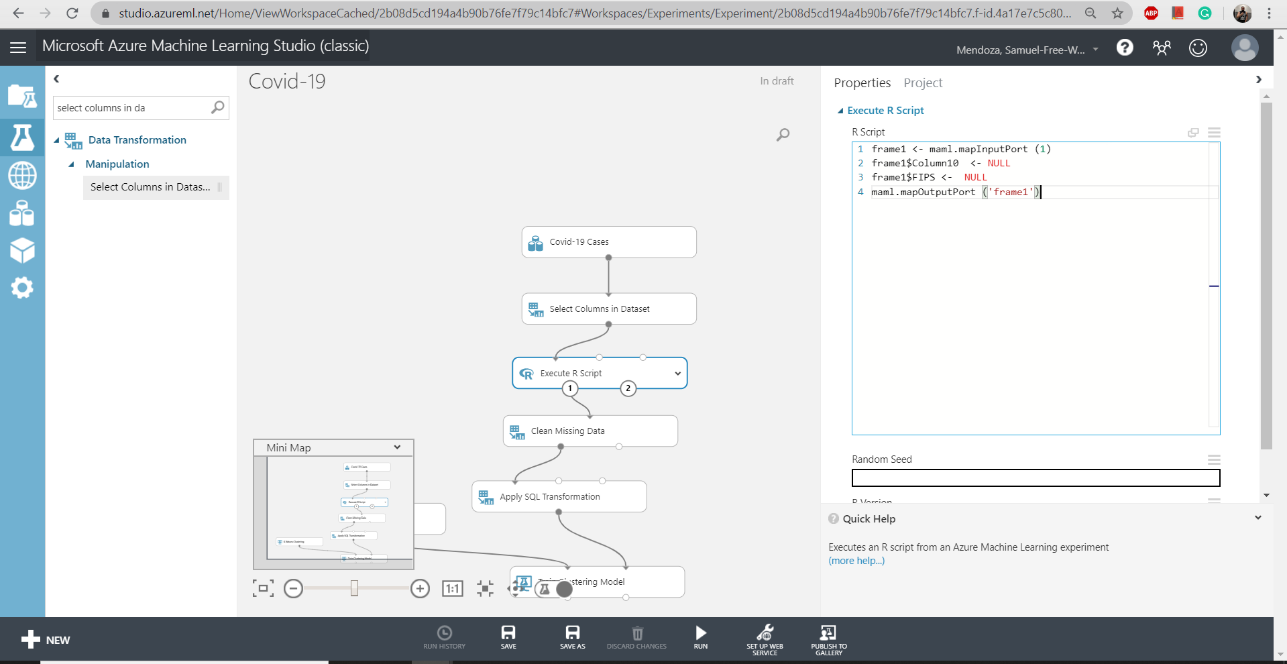
**This step allows use to remove columns that are not necessary to the project to better optimize the data.**

1. In the search box of the experiment pane, type **Select Columns in Dataset**, drag Select Columns in Dataset under Covid-19 Cases. Click the output port of Covid-19 Cases and drag it to the input of Select Columns in Dataset.
2. Click on **Select Columns in Dataset,** on the properties pane, click on **Launch Column Selector**
3. **With Rules,** click **All Columns** and **Exclude Column Names:** Admin2, Column 11, Column12, Column 13, Column 14, Column 15, Column 16, Column 17, Column 18, and Column 10



Step 4: Execute R Script

**The R Script allows us to remove the FIPS Column to better optimize the dataset.**

1. In the search box of the experiment pane, type **Execute R Script**, drag Execute R Script under Select Columns in Dataset. Click the output port of Select Columns in Dataset and connect it to leftist input of Execute R Script.
2. Download and open the Azure code from the GitHub. Copy the Execute R Script. Back in the studio, go to the Execute R Script properties pane. Select the current code in R script and paste the code from the Azure code file.

**EXECUTE R SCRIPT:**

frame1 <- maml.mapInputPort (1)

frame1$Column10 <- NULL

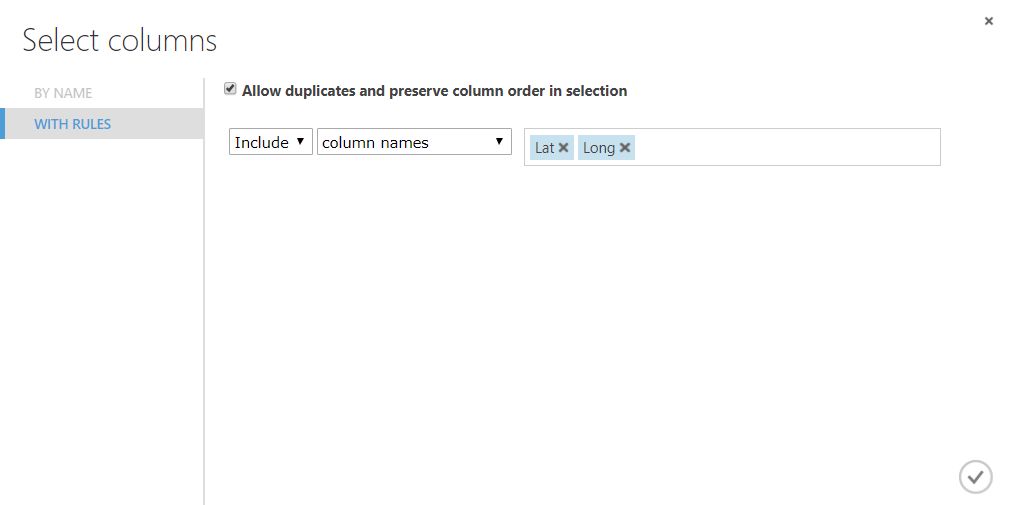
frame1$FIPS <- NULL

maml.mapOutputPort ('frame1')

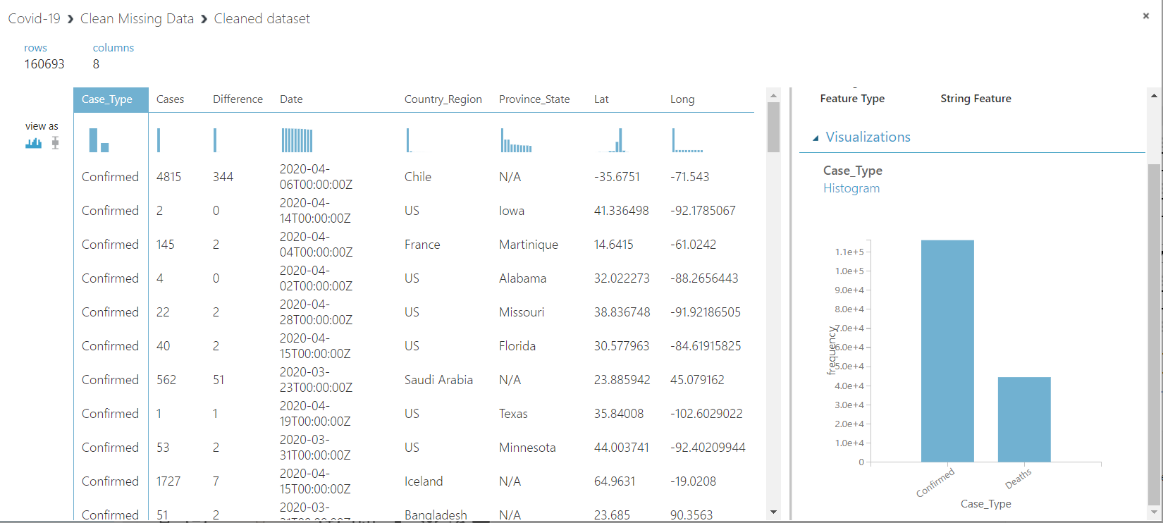
Step 5: Clean Missing Data

**Using the Clean Missing Data allows us to remove the missing information from Latitude and Longitude to better optimize these columns for visualization.**

1. In the search box of the experiment pane, type **Clean Missing Data** and drag it under Execute R Script. Click the left-output port of Execute R Script and connect it to input of Clean Missing Data.
2. In the properties pane of Clean Missing Data, Launch Column Selector. With Rules Include Column Names LAT and LONG. Check Allow duplicates and preserve column order in selection.



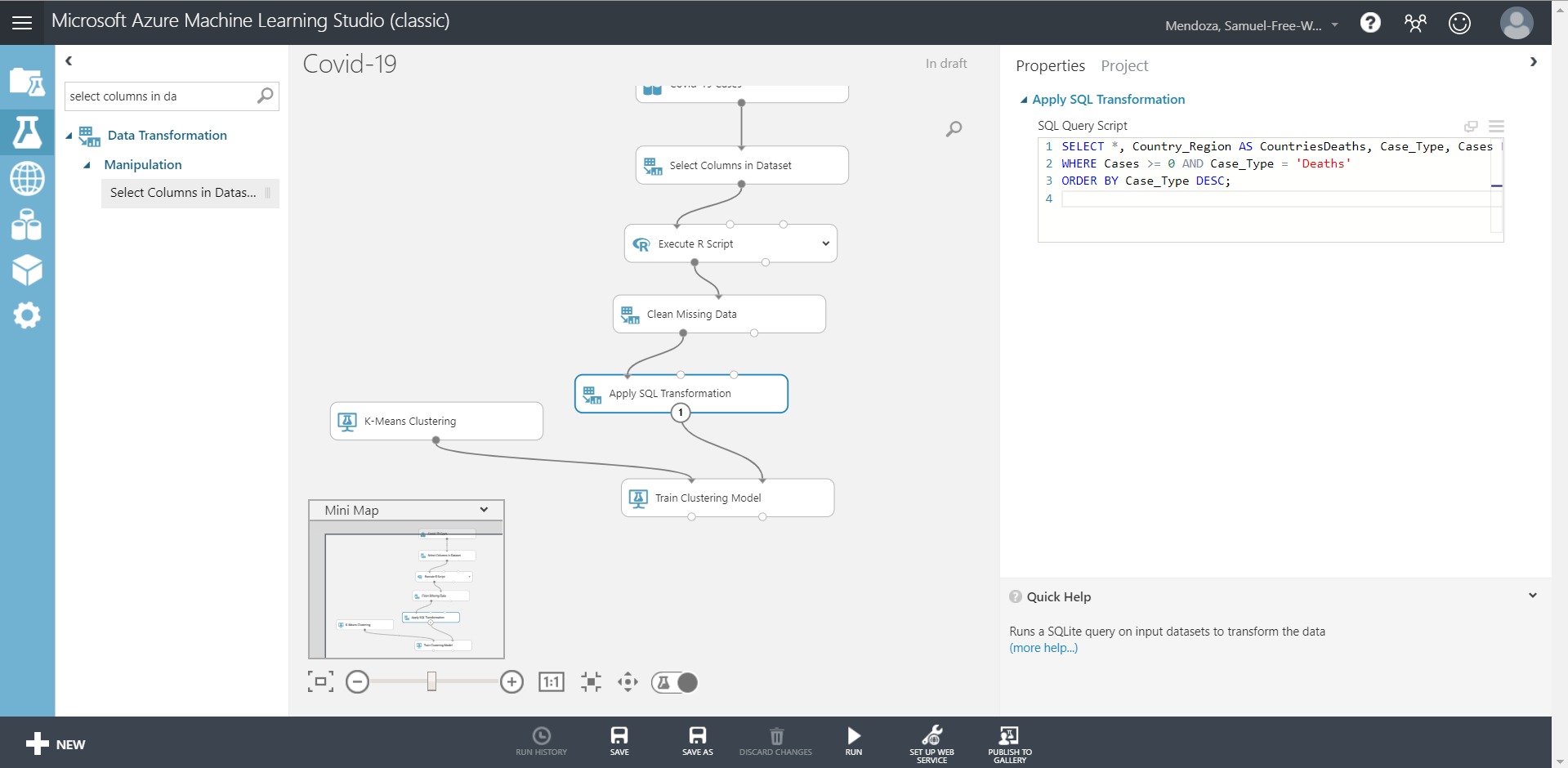
1. In the Properties Pane:
   * **Minimum missing value ratio:** 0
   * **Maximum missing value ratio:** 1
   * **Cleaning Mode:** Custom Substitution Value
   * **Replacement Value:** 0
   * **Uncheck Generate missing value indicator column.**
2. **Save and Run**
3. After it is finished running, look for the **Green checkmark,** Visualize the **Cleaned dataset** to see the histogram of Case\_Types



Step 6: Apply SQL Transformation

**Using the SQL Query Script, we query the dataset by creating a view for the death count instead of viewing the both death and confirmed cases.**

1. In the search box of the experiment pane, type **Apply SQL Transformation** and drag it under Clean Missing Data. Click the left-output port of Clean Missing Data and connect it to the left input of Apply SQL Transformation.
2. Open the Azure code from the Github, Copy the Apply SQL Transformation script. Back in the studio, go to the Apply SQL Transformation properties pane, click in the SQL Query Script, Select the current script and paste the new SQL script.



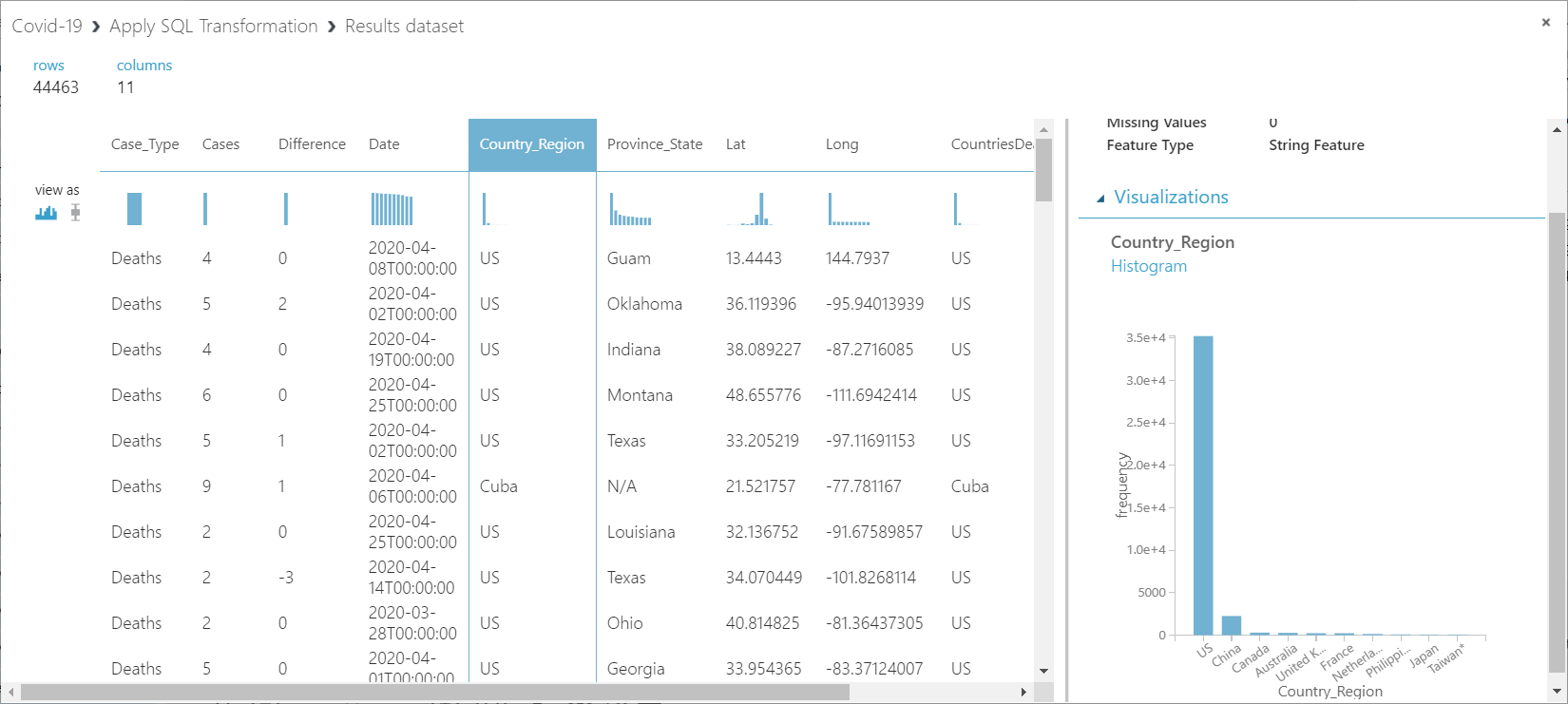
**APPLY SQL TRANSFORMATION:**

SELECT \*, Country\_Region, Cases, Case\_Type FROM t1

WHERE Cases >= 0 AND Case\_Type = 'Deaths'

ORDER BY Case\_Type DESC;

1. **Save and Run**
2. After it is finished running, look for the **Green checkmark,** Visualize the **Result dataset** of the SQL Transformation. Here you will be able to see more statistics plus the A Histogram on the Column County\_Region where you can see the comparison between the United States and some of the other nations from around the world.



Step 7: K-Means Clustering

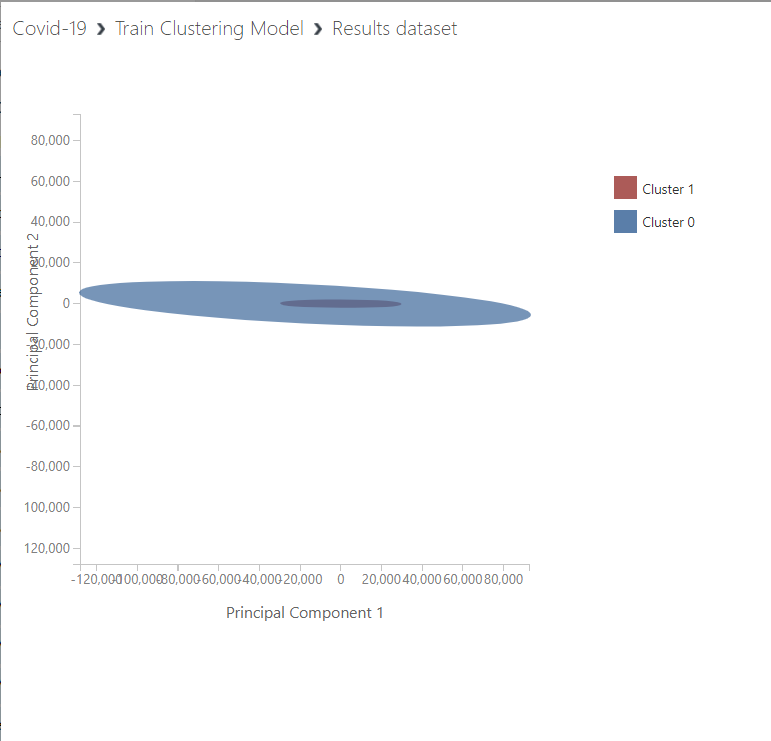
**K-Means Clustering allows us to group together certain parts of the data and randomly initialize center in hopes to prepare data for a trained model algorithm.**

1. In the search box of the experiment pane, type **K-Means Clustering** and drag it to the left of Apply SQL Transformation.
2. In the K-Means Clustering properties pane apply the following:
   * **Create trainer mode:** Single Parameter
   * **Number of Centroids:** 2
   * **Initialization:** Random
   * **Random number seed:** 2345
   * **Metric:** Euclidean
   * **Iteration:** 100
   * **Assign Label Mode:** Fill Missing Values

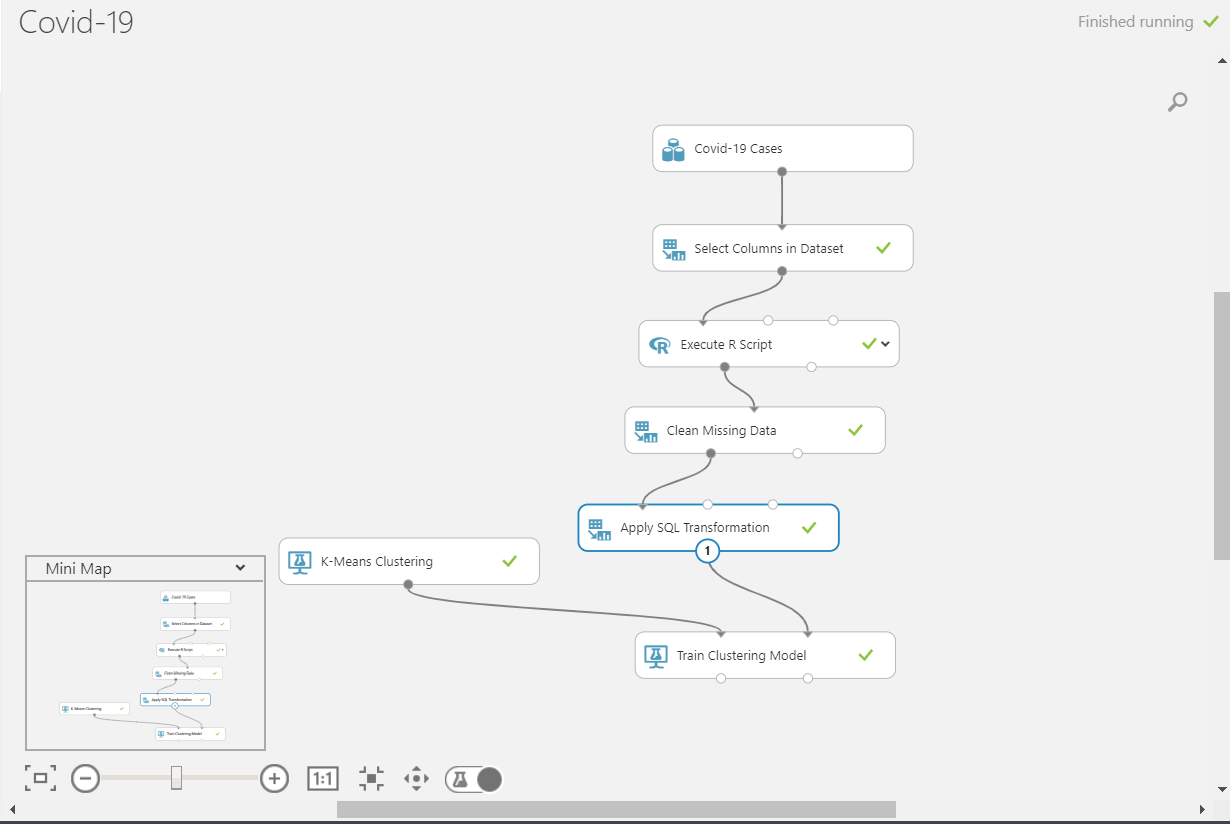
Step 8: Train Clustering Model

**Taking the previous step in clustering the data, we train it to output a visualization to compare the scope of confirmed cases to deaths.**

1. In the search box of the experiment pane, type **Train Clustering Model** and drag it under Apply SQL Transformation. Click the output port of Apply SQL Transformation and connect it to the right input of Train Clustering Model.
2. Connect **K-Means Clustering** Output port into **Train Clustering Model** left input.
3. In the Train Clustering Model properties pane, launch column selector.
4. With Rules, No Columns 🡪 Include Column names 🡪Case\_Type, Difference, Cases, CountriesDeaths. Apply changes.
5. Check for Append or Uncheck for Result Only.
6. **Save and Run**
7. After it is finished running, look for the **Green checkmark,** Visualize **Result Dataset** of the Train Clustering Model. You will see a comparison between confirmed and death from Covid-19 Virus.



1. By the end, your Canvas should look like this:



References

* 1. Data World, John Hopkins University Dataset, <https://data.world/covid-19-data-resource-hub/covid-19-case-counts/workspace/file?filename=COVID-19+Cases.csv>
  2. GitHub, https://github.com/juliom12/covid\_19\_cases
  3. Dropbox, <https://www.dropbox.com/s/smw5aein7i8zr1d/covid-19-data-resource-hub-covid-19-case-counts%20%281%29.zip?dl=0>
  4. Microsoft Azure ML, <https://studio.azureml.net/>